The Journal of Open Source Software

{fairmetrics}: An R package for group fairness evaluation

Benjamin Smith[1], **Jianhui Gao**[1], **and Jessica Gronsbell**[1]

1 Department of Statistcial Science, University of Toronto

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

Fairness is a growing area of machine learning (ML) that focuses on ensuring models do not produce systematically biased outcomes for particular groups, particularly those defined by protected attributes such as race, gender, or age. Evaluating fairness is a critical aspect of ML model development, as biased models can perpetuate structural inequalities. The {fairmetrics} R package offers a user-friendly framework for rigorously evaluating numerous group-based fairness criteria, including metrics based on independence (e.g., statistical parity), separation (e.g., equalized odds), and sufficiency (e.g., predictive parity). Group-based fairness criteria assess whether a model is equally accurate or well-calibrated across a set of predefined groups so that appropriate bias mitigation strategies can be implemented. {fairmetrics} provides both point and interval estimates for multiple metrics through convenient wrapper functions and includes an example dataset derived from the Medical Information Mart for Intensive Care, version II (MIMIC-II) database (Goldberger et al. 2000; J. Raffa 2016).

Statement of Need

ML models are increasingly integrated into high-stakes domains to support decision making that significantly impacts individuals and society more broadly, including criminal justice, healthcare, finance, employment, and education(Mehrabi et al. 2021). Mounting evidence suggest that these models often exhibit bias across groups defined by protected attributes. For example, within criminal justice, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software, a tool used by U.S. courts to evaluate the risk of defendants becoming recidivists, was found to incorrectly classify Black defendants as high-risk at nearly twice the rate of white defendants(Mattu n.d.). This bias impacted Black defendants by potentially leading to harsher bail decisions, longer sentences, and reduced parole opportunities compared to white defendants with similar risk profiles. Similarly, within healthcare, a commercial risk-prediction algorithm deployed in the U.S. to identify patients with complex health needs for high-risk care management programs was shown to be significantly lower calibrated for Black patients relative to white patients(Obermeyer et al. 2019). This caused Black patients with equivalent health conditions to be under-referred for essential care services compared to white patients. These examples illustrate that there is an urgent need for ML practitioners and researchers to ensure that ML models support fair decision before they are deployed in real-world applications.

While existing software can compute group fairness criteria, they only point estimates and/or visualizations without quantifying the uncertainty around the criteria. This limitation prevents users from determining whether observed disparities between groups are statistically significant or merely the result of random variation due to finite sample size, potentially leading to incorrect conclusions about fairness violations. The {fairmetrics} R package addresses this gap by providing bootstrap-based confidence intervals (CIs) for both difference-based and ratio-based group fairness metrics, empowering users to make statistically grounded decisions about the fairness of their models, which is inconsistently done in practice.

Scope

The `{fairmetrics}` package is designed to evaluate group fairness in the setting of binary classification with a binary protected attribute. This restriction reflects standard practice in the fairness literature and is motivated by several considerations. First, binary classification remains prevalent in many high-stakes applications, such as loan approval, hiring decisions, and disease screening, where outcomes are typically framed as accept/reject or positive/negative(Mehrabi et al. 2021). Second, group fairness is the most widely used framework for binary classification tasks(Mehrabi et al. 2021). Third, when protected attributes have more than two categories, there is no clear consensus on how to evaluate group fairness. This focus enables {fairmetrics} to provide statistically grounded uncertainty quantification for group fairness metrics commonly applied in binary classification tasks across diverse application domains.

Fairness Criteria

Group fairness criteria are typically classified into three main categories: independence, separation, and sufficiency (Barocas, Hardt, and Narayanan 2023; Berk et al. 2018; Castelnovo et al. 2022). Independence requires that the model's predictions be statistically independent of the protected attribute, meaning the likelihood of receiving a positive prediction is the same across protected groups. Separation requires independence between the prediction and the protected attribute conditional on the true outcome, so that the probability of a positive prediction is equal across protected groups within positive (or negative) outcome class. Sufficiency requires independence between the outcome and the protected attribute conditional on the prediction, implying that once the model's prediction is known, the protected attribute provides no additional information about the true outcome.

Independence

- **Statistical Parity:** Compares the overall rate of positive predictions between groups, irrespective of the true outcome.
- **Conditional Statistical Parity:** Restricts the comparison of positive prediction rates to a specific subgroup (e.g., within a hospital unit or age bracket), offering a more context-specific fairness assessment.

Separation

- **Equal Opportunity:** Compares disparities in false negative rates between groups, quantifying any difference in missed positive cases.

- **Predictive Equality:** Compares false positive rates (FPR) between groups, ensuring that no group is disproportionately flagged as positive when the true outcome is negative.
- **Balance for Positive Class:** Compares if the average of predicted probabilities among individuals whose true outcome is positive is similar between groups.
- **Balance for Negative Class:** Compares if the average of predicted probabilities among individuals whose true outcome is negative is similar between groups.

Sufficiency

- **Predictive Parity:** Compares positive predictive values across groups, assessing whether the precision of positive predictions is equivalent.

Other Criteria

- **Brier Score Parity:** Compares the Brier score—the mean squared error of predicted probabilities—is similar across groups, indicating comparable calibration.
- **Accuracy Parity:** Compares the overall accuracy of a predictive model is equivalent across groups.
- **Treatment Equality:** Compares the ratio of false negatives to false positives across groups, ensuring the balance of missed detections versus false alarms is consistent.

Evaluating Fairness Criteria

The primary input to the {fairmetrics} package is a data frame or tibble which containing the model's predictions, true outcomes, and the protected attribute in question. Figure 1 shows the workflow for using {fairmetrics}. It is possible to evaluate a model for a specific or multiple group fairness metrics.

A simple example of how to use the {fairmetrics} package is shown below. The example makes use of the `mimic_preprocessed` dataset, a pre-processed version of the the Indwelling Arterial Catheter (IAC) Clinical Dataset, from MIMIC-II clinical database[1] (J. Raffa 2016; J. D. Raffa et al. 2016). This dataset consists of 1776 hemodynamically stable patients with respiratory failure, and includes demographic information (patient age and gender), vital signs, laboratory results, whether an IAC was used, and a binary outcome indicating whether the patient died within 28 days of admission.

While the choice of fairness metric used is context dependent, we show all metrics available with the `get_fairness_metrics()` function. In this example, we evaluate the model's fairness with respect to the protected attribute `gender`. For conditional statistical parity, we condition on patients older than 60 years old. The model is trained on a subset of the data, and predictions are made on a test set.

```
library(fairmetrics)
library(dplyr)
library(magrittr)
library(randomForest)
```

[1]The raw version of this data is made available by PhysioNet (Goldberger et al. 2000) and can be accessed in {fairmetrics} package by loading the `mimic` dataset.

The Journal of Open Source Software

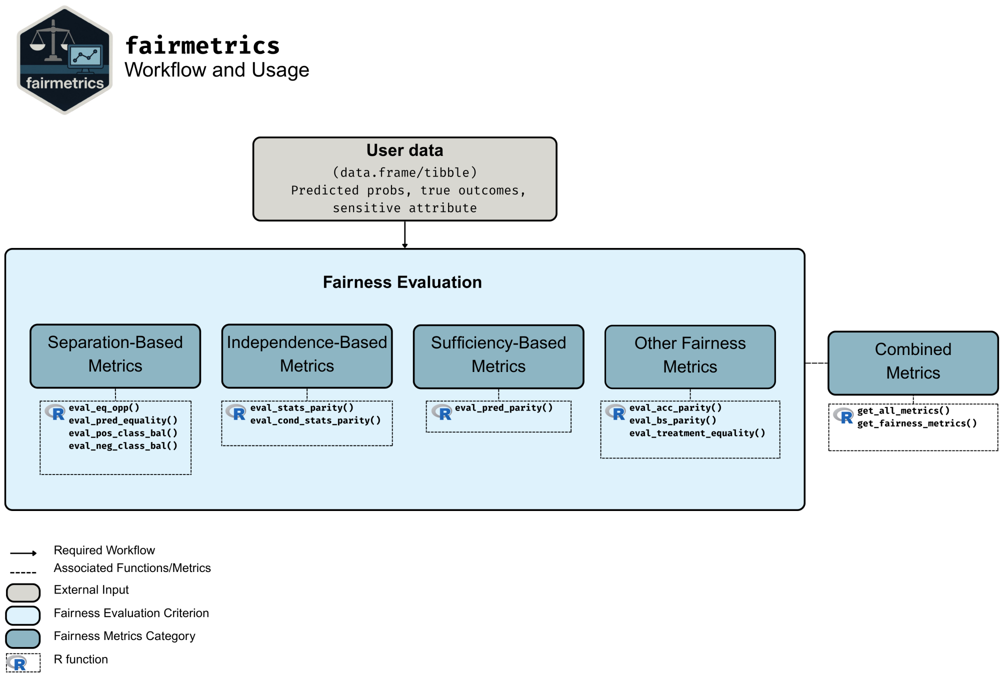

Figure 1: Workflow for using {fairmetrics} to evaluate model fairness across multiple criteria.

```r
# Load the example dataset
data("mimic_preprocessed")

# Split the data into training and test sets
train_data <- mimic_preprocessed %>%
  dplyr::filter(dplyr::row_number() <= 700)

test_data <- mimic_preprocessed %>%
  dplyr::mutate(gender = ifelse(gender_num == 1, "Male", "Female")) %>%
  dplyr::filter(dplyr::row_number() > 700)

# Train a random forest model
rf_model <- randomForest::randomForest(
  factor(day_28_flg) ~ .,
  data = train_data,
  ntree = 1000
  )

# Make predictions on the test set
test_data$pred <- predict(rf_model, newdata = test_data, type = "prob")

# Evaluate predictive equality
# (Setting message=FALSE to avoid cluttering the output)

get_fairness_metrics(
 data = test_data,
 outcome = "day_28_flg",
```

```
  group = "gender",
  group2 = "age",
  condition = ">=60",
  probs = "pred",
  cutoff = 0.41
)

#>                                         Metric GroupFemale GroupMale Difference
#> 1                           Statistical Parity        0.17      0.08       0.09
#> 2  Conditional Statistical Parity (age >=60)          0.34      0.21       0.13
#> 3                            Equal Opportunity        0.38      0.62      -0.24 [
#> 4                          Predictive Equality        0.08      0.03       0.05
#> 5                   Balance for Positive Class        0.46      0.37       0.09
#> 6                   Balance for Negative Class        0.15      0.10       0.05
#> 7                            Predictive Parity        0.62      0.66      -0.04
#> 8                           Brier Score Parity        0.09      0.08       0.01
#> 9                      Overall Accuracy Parity        0.87      0.88      -0.01
#> 10                          Treatment Equality        1.03      3.24      -2.21 [
```

Among the fairness metrics calculated, metrics whose difference confidence intervals cross zero and whose ratio confidence intervals cross one indicate no significant difference between the groups. In this example, the statistical parity, conditional statistical pairty, equal opportunity, predictive equality, positive class balance, negative class balance, and treatment equaltiy metrics show a significant differences. While, the predictive parity, Brier score parity, and overall accuracy parity metrics do not show significant differences between the groups.

Should the user wish to calculate an individual metric, it is possible to use any of the `eval_*` functions. For example, to calculate the equal opportunity metric, the user can use the `eval_equal_opportunity()` function.

```
eval_eq_opp(
  data = test_data,
  outcome = "day_28_flg",
  group = "gender",
  probs = "pred",
  cutoff = 0.41
)

#>There is evidence that model does not satisfy equal opportunity.
#>  Metric GroupFemale GroupMale Difference    95% Diff CI Ratio 95% Ratio CI
#>1    FNR        0.38      0.62      -0.24 [-0.39, -0.09]  0.61 [0.44, 0.85]
```

For group-specific performance metrics, such as the True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), Negative Predictive Value (NPV), and others, the `get_all_metrics()` function can be used. This function returns a data frame with the metrics calculated for each group.

```
get_all_metrics(
  dat = test_data,
  outcome = "day_28_flg",
  group = "gender",
  probs = "pred",
  cutoff = 0.41
```

```
)

#>          Metric Group Female Group Male
#> 1          TPR         0.62       0.38
#> 2          FPR         0.08       0.03
#> 3          PPR         0.17       0.08
#> 4          PPV         0.62       0.66
#> 5          NPV         0.92       0.90
#> 6          ACC         0.87       0.88
#> 7   Brier Score        0.09       0.08
#> 8   FN/FP Ratio        1.03       3.24
#> 9 Avg Pred Prob        0.21       0.14
```

Related Work

Other R packages similar to {fairmetrics} include {fairness}(Kozodoi and V. Varga 2021), {fairmodels} (Wiśniewski and Biecek 2022) and {mlr3fairness}[mlr3fairness_package]. The differences between {fairmetrics} and these other packages is twofold. The primary difference between is that {fairmetrics} calculates the ratio and difference between group fairness criterion and allows estimated confidence intervals of fairness metrics via bootstrap - allowing for more meaningful inferences about the fairness metrics calculated. Additionally, in contrast to the {fairmodels}, {fairness} and {mlr3fairness} packages, the {fairmetrics} package does not posses any external dependencies and has a lower memory footprint, resulting in an environment agnostic tool that can be used with modest hardware and older systems. Table 1 shows the comparison of memory used and dependencies required when loading each library.

Package	Memory (MB)	Dependencies
fairmodels	17.02	29
fairness	117.61	141
fairmodels	58.11	45
fairmetrics	0.05	0

Table 1: Memory usage and dependencies of fairmetrics vs similar packages (MB)

For python users, the {fairlearn} library (Weerts et al. 2023) provides a broader set of fairness metrics and algorithms. The {fairmetrics} package is designed for seemless integration with R workflows, making it a more convenient choice for R-based ML applications.

Licensing and Availability

The {fairmetrics} package is under the MIT license. It is available on CRAN and can be installed by using `install.packages("fairmetrics")`. A more in-depth tutorial can be accessed at: https://jianhuig.github.io/fairmetrics/articles/fairmetrics.html. All code is open-source and hosted on GitHub. All bugs and inquiries can be reported at https://github.com/jianhuig/fairmetrics/issues/.

References

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, Massachusetts: The MIT Press.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research* 50 (1): 3–44. https://doi.org/10.1177/0049124118782533.

Castelnovo, Alessandro, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. "A Clarification of the Nuances in the Fairness Metrics Landscape." *Scientific Reports* 12 (1). https://doi.org/10.1038/s41598-022-07939-1.

Goldberger, Ary L., Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." *Circulation [Online]* 101 (23): e215–20. https://doi.org/10.1161/01.CIR.101.23.e215.

Kozodoi, Nikita, and Tibor V. Varga. 2021. *Fairness: Algorithmic Fairness Metrics*. https://CRAN.R-project.org/package=fairness.

Mattu, Lauren Kirchner, Jeff Larson. n.d. "Machine Bias." *ProPublica*. https://www.propublica.org bias-risk-assessments-in-criminal-sentencing. Accessed May 29, 2025.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Comput. Surv.* 54 (6). https://doi.org/10.1145/3457607.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53.

Raffa, Jesse. 2016. "Clinical Data from the MIMIC-II Database for a Case Study on Indwelling Arterial Catheters (Version 1.0)." https://doi.org/10.13026/C2NC7F. https://doi.org/10.13026/C2NC7F.

Raffa, Jesse D., Mohammad Ghassemi, Tristan Naumann, Mengling Feng, and Daniel J. Hsu. 2016. "Data Analysis." In *Secondary Analysis of Electronic Health Records*, 109–22. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_9.

Weerts, Hilde, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. "FairLearn: Assessing and Improving Fairness of AI Systems." *arXiv.org*. https://arxiv.org/abs/2303.16626.

Wiśniewski, Jakub, and Przemysław Biecek. 2022. "Fairmodels: A Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models." *The R Journal* 14 (1): 227–43. https://doi.org/10.32614/RJ-2022-019.